GESTURAL CONTROL OF REAL-TIME SPEECH SYNTHESIS IN LUNA PARK

Grégory Beller
IRCAM, Paris
beller@ircam.fr

ABSTRACT

This paper presented the researches and the developments realized for an artistic project called *Luna Park*. This work is widely connected, at various levels, in the paradigm of the concatenative synthesis, both to its shape and in the processes which it employs. Thanks to a real-time programming environment, synthesis engines and prosodic transformations are manipulated, controlled and activated by the gesture, via accelerometers realized for the piece. This paper explains the sensors, the real time audio engines and the mapping that connects this two parts. The world premiere of *Luna Park* takes place in Paris, in the space of projection of the IRCAM, on June 10th, 2011, during the festival AGORA.

1. INTRODUCTION

Luna Park is a piece of musical theater, of duration about one hour, written by Georges Aperghis, staged by Daniel Levy, and whose computer music is designed by Grégory Beller. The general subject of the piece approaches the way the electronic surveillance and the massive collection of personal digital data make of our current world, a gigantic park of attraction. Four performers are Eva Furrer, octabass flute and voice, Johanne Saunier, dance and voice, Mike Schmidt, bass flute and voice, and Richard Dubelsky, air percussions and voice. The fact that they have all four vocal parts to perform, as well as the scenography of the set (including video), move *Luna Park* closer to a previous work of G. Aperghis, called Machinations (2002). However, this new work distinguishes itself from the previous one, notably by gesture sensors' use and from the speech synthesis. Indeed, various sensors, (accelerometers, tactile ribbons and piezoelectric sensors) were developed and realized to allow the performers to control various audio engines, by the gesture. The mapping between the data stemming from these sensors and the various audio processings, realized within the real-time programming environment Max/MSP, is different from a sequence in the other one and can evolve in the time.

That is why this article is presented according to the following plan. In a first part, this article lists the various sensors realized for this creation and gives details of their developments, as well as the data which they produce. In a second part, the realized audio engines are described under the shape of real-time processes. Besides the concatenative synthesis engine, is presented an innovative engine of prosodic transformation allowing the real time modification of the speech rate. The third part proposes some examples of mapping between the sensors data and the audio engines parameters, notably used for the piece. Finally, the fourth part allows to conclude and to propose some perspectives.

2. GESTURE CAPTURE

2.1 Background

In the case of the acoustic instruments, a gesture is necessary for the sound production. This is not any more the case of the electronic instruments where the control of electronic sound processes is separated from the process of sound production. Several projects of research and creation in the IRCAM use the gesture capture. Whether it is to augment acoustic instruments, (Bogen Lied [1], Augmented Violin Project [2], augmented percussions of Fedele) or to connect different worlds together such as music, video, gesture and dance for the creation (Glossopoeia) or for the pedagogy [3], certain number of sensors was realized in the IRCAM. However, the gesture capture was neither used yet to control prosodic modifications nor vocal synthesizers in the spectrum of the contemporary creation in the IRCAM. It is thus a new way to use the gesture capture systems that we propose for this project of research and creation. The gestural control of the speech synthesis constitutes henceforth a complete field of research. Controllers of various types were elaborated for various types of synthesizers of spoken voice, or sung voice [4]. Among these mappings, we find the "Speech Conductor" [5, 6], the "Glove-Talk" [7], the "Squeeze Vox" [8], the "SPASM" [9], the "OUISPER" project [10, 11] and some others [12]. We chose to use the movements of the hand for several reasons, besides crossing the scenographic reasons. First of all, the spontaneous speech can be naturally accompanied with a movement of hands. The idea to accompany the movements of hands by the speech, by the reversibility, seems thus natural. The percussive aspect of the movements fits the concatenative synthesis in which segments are activated in a discrete way in the time, so managing the segmental aspect of the speech. On the contrary, the continuous aspect of the movements of hands allows a control of the prosody, the suprasegmental aspect of

the speech. If we consider the classic asymmetry right-left such as know it the conductors (for the right-handers, the left hand is rather connected with the expression, whereas the right hand is rather connected with the important temporal markers of the music), we can then create a gestural control of both hands of the synthesis, with for a right-hander, a right hand managing the segmental aspect and a left hand managing the suprasegmental aspect. It is one of the possible scenarios that we exploited for the creation of *Luna Park* (see section 4).

2.2 Accelerometers gloves

The technology of gloves wireless accelerometers / gyroscopes used [13], presented on figure 1 allows to measure the accelerations of both hands according to 6 axes (3 in translation and 3 in rotation with gyroscopes). The raw data delivered by gloves are not necessarily easy to interpret. So a first stage of preprocessing allows to return more interpretable data.

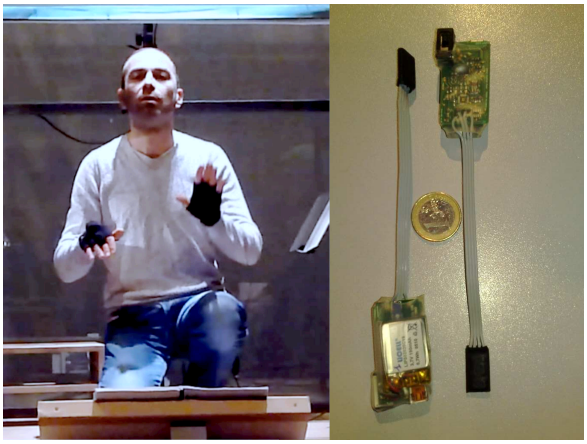

Figure 1. Images of the sensors (to the right) and of Richard Dubelsky who wears them in hands thanks to gloves (to the left).

2.2.1 Preprocessing

The data resulting from the wifi receiver are transmitted via UDP every 1 ms. To synchronize them to the internal clock of Max/MSP, they are first median filtered (order 5) and sub-sampled by a factor 5. Thus we obtain a stable stream of data every 5 ms. Then various descriptors of the gesture arise from these preprocessed raw data.

2.2.2 Variation of the momentum

The estimate of the immediate acceleration allows to know, at any time, the variation of momentum relative to the gesture. This momentum, according to the laws of the classical mechanics, is directly proportional in the speed. The raw data coming from the sensor are at first “denoised” thanks to the average on the last 4 samples. The root of the sum of the square of these six filtered values allows to obtain a proportional quantity in the variation of momentum of the gesture.

2.2.3 Hit energy estimation

The hit energy estimation allows the immediate release from the observation of the variation of the momentum of the gesture. Three values, delivered by the sensors of acceleration in translation, are stored in a circular buffer including all the time, 20 samples. Three standard deviation corresponding to these values are added, all the time, (norm I corresponding to the sum of the absolute values). This sum also allows to represent the variation of momentum of the gesture. To detect variation of this value, corresponding to abrupt variations of the gesture, it is compared all the time with its median value (order 5). When the difference between these two values exceed certain arbitrary threshold, a discrete value appears to mean the presence of a fast change of the gesture. It allows, for example, to emit a regular click, when we beat a measure with the hand, every time the hand changes direction. The hit energy estimation is a process allowing to generate discrete data from a gesture, by definition continuous. Indeed, of a continuous physical signal, it allows by thresholding, to define moments corresponding to the peaks of variation of the momentum, which coincide, from a perceptive point of view for the user, in peaks of efforts (of acceleration). By this process, it becomes then possible to create precise air percussions either sounds activation at the moment when the hand of the user changes direction or accelerates surreptitiously.

2.2.4 Absolute position of the hand

The Earth´s gravitational field introduces an offset into the answer of the sensors which can be exploited to deduct the absolute position of the hands, as well as the presence of slow movements. This quasi-static measure brings a continuous controller to the performer. A playful example of the use of this type of data is the air rotary potentiometer in which the rotation of the hand can control the volume (or other) of a sound.

2.3 Piezoelectric sensors

Besides the accelerometers gloves, the percussionist plays physical percussions. He emits sounds by striking certain zones of his body and/or the structure surrounding him. The integration of local and sensitive zones to the touch, on a suit, is not easy. The traditional pads of an electronic drum kit are too stiff and non-adapted to hand strikings. We chose to use smaller and more flexible, piezoelectric microphones. Two microphones (one near the left hip and the other one near the right shoulder) placed on the percussionist allow him to play with two different zones of his body. Six microphones of the same type are also arranged in its surrounding space. The audio signals delivered by these various microphones are processed by a classical attack detection system allowing the percussionist to activate various types of sounds according to zones. Likely if the

pads of an electronic drum kit was arranged on and around the percussionist.

3. AUDIO ENGINES

3.1 Real time concatenative synthesis

The concatenative synthesis is realized in real time thanks to the object Mubu.concat developed in association with the team Real-Time Musical Interaction of the IRCAM [14, 15]. This object contains many functionalities of the cataRT patch [16]. The object takes, as input, a sound file (buffer) and an associated markers file. He allows the reading of segments by the choice of their indexes. It also includes other options affecting the reading, such as the transposition, the windowing, or still the used output. Segments can succeed one another automatically or be launched by a metronome or another discrete signal emitted by a sensor (stemming from the hit energy estimation, for example). The order of segments is arbitrary and all the sequences of index are possible. Among them, an incremental series of step 1 will restore the original audio file without audible presence of the segmentation, whereas a random series will generate a variation of the starting material and will make audible the beforehand chosen segmentation. Various methods to generate interesting sequences of index within the framework of the speech synthesis are presented below.

3.2 Speech synthesis

If the audio engine of concatenative synthesis does not need to be sophisticated, compared with the other paradigms of synthesis, such as the HMM-based speech synthesis or still the articulatory synthesis, it is because the intelligence of the concatenative speech synthesis rests on the selection of segments, that is, on the definition of the sequence of the indexes. Indeed, the concatenative speech synthesizer bases on two distances allowing to define simultaneously the closeness of the result with regard to a target (distance of the target) and the quality of this result (distance of concatenation).

3.2.1 Distance of the target

The first distance, as its name indicates it, requires the definition of a target. In the Text-To-Speech synthesis, this target is defined in a symbolic way by the text and by the various analyses which derive from it (grammar, phonetics...). In the hybrid synthesis speech/music [17], a target can be defined in a acoustic way as a sequence of audio descriptors. Any targets may be used as long as it shares with segments, a common descriptor. The synthesizer IrcamTTS [18, 19], presents a peculiarity face to face the other TTS (Text-To-Speech) synthesizers, because he allows the user to define his target in a symbolic way, by the text, but also in a acoustic way, by some prosodic symbols. So the user can write on the same support, in a joint way, the wished text and the way he would like this text is pronounced. This tool is very appreciated, consequently, by composers who can write not only the text, but also the prosody they wish, like a score. The target can be also defined in real time.

3.2.2 Distance of concatenation

The distance of concatenation allows to estimate the perceptive weight of the concatenation of two segments. Naturally consecutive segments cause a zero weight. Segments from which spectrum in edges are very different, will produce, a higher weight, supposed to mean the introduction of an synthesis artifact.

3.2.3 Speech synthesis in batch mode

As the definition of a real time target is not a common matter, most of the TTS synthesizers work in batch mode. The user writes a sentence, then chooses generally a voice, to pronounce it. The most spread selection algorithm of segments then appeals a Viterbi decoding allowing the joint minimization of the target distance and the distance of concatenation on the whole sentence to synthesize. The "backward" phase of this algorithm allowing the definition of the optimal solution requires to know the end of the sentence to select the beginning, what makes the synthesis engine profoundly not real-time.

3.3 Real time speech synthesis

A real time speech synthesizer has sense only if the text is generated in real time. The case appears (as shawn in section 4) when the text is generated by statistical models (HMM, N-gram, K-NN) which transforms or generates one text on the fly. The real-time constraint does not allow us any more to use the phase "backward" of the Viterbi algorithm guaranteeing the optimality of the path on a whole sentence, because we do not know early, the end of the current sentence. The joint minimization of the distance in the target and the distance of concatenation can be made then only between every segment, in a local way and not in a global one. The advantage of this method lies in its ability to react, while its inconvenience is that it produces a sound result poorer than the classic batch TTS synthesis. In *Luna Park*, several paradigms are used to generate, in real time, targets which guide the concatenative synthesis.

3.3.1 Predefined sequences

First of all, certain texts are fixed, a priori, and modeled under the shape of a sequence of words, syllables, phones or semi-phones. Segments constituting these sequences are then launched by the performers and can be chosen according to data stemming from sensors or from vocal analyses (see section 4). It allows for expected syllables series, and to produce a clear semantic sense.

3.3.2 Predefined sequences and N-gram

It happens that several segments possess the same symbol (several phones corresponding to the same phoneme, for example). We can then estimate the probability of transition from a symbol to the other one and make random series respecting more or less the initial sequence. Thanks to N-gram of order N variable, we can control in real time the closeness of the generated sequence with regard to the

predefined text (the bigger N, the closer to the initial sequence; the smaller N, the further to the initial sequence). It notably allows the performers to control the semantic aspect of the output.

3.3.3 Predefined sets and HMM

Like a text, some segment sets (syllables, phones, words) were also predefined. In the same way, segments belonging to these sets are activated by the performers according to descriptors or in a random way. It notably allows to create textures (of phones or syllables) or still rhythms from a single syllable. An interface was create to allow to choose in real time, the probability of appearance of such symbols. It appears under the shape of a histogram of the available symbols. The modification of this histogram allows to modify the probability of transition from a symbol to the other one (HMM of order 1). Once the symbol is chosen, a corresponding segment can be activated according to various controllers / descriptors values. The figure 2 presents the interface that permits to generate in real time targets. At the top, a recording segmented in syllables thanks to the program IrcamAlign [20] which allows the speech segmentation in various units in batch mode), is the input. In the middle, the histogram presenting the relative rate of appearance of their symbols. Below, the same modified histogram allows to modify the probability of appearance of the activated syllables. For example, the generated output is composed of an equiprobable set of syllables “ni”, “naR” and “noR”.

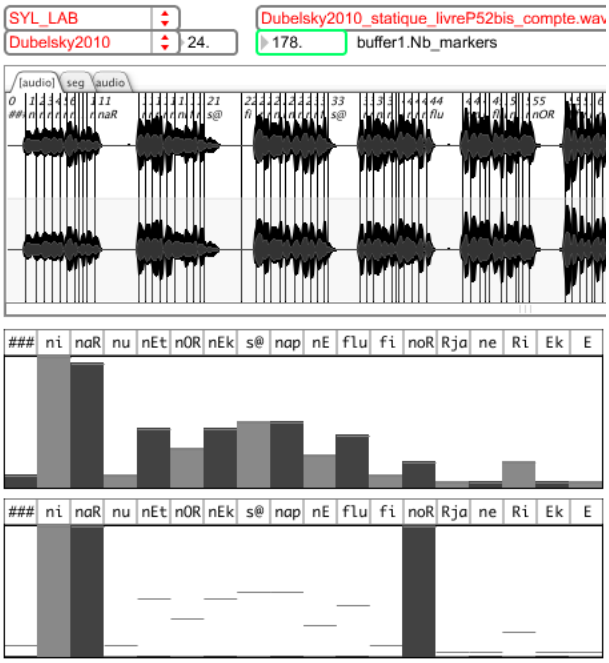

Figure 2. Interface for the definition of the real time target by editing histogram.

3.4 Real time prosodic transformation

Real time prosodic transformation allows many applications [21]. It is possible thanks to speech analysis/synthesis paradigm allowing to estimate and to modify the prosodic dimensions of the speech. The prosodic transformation can apply to the voice of the performers or to the synthesis output in a equivalent way. The prosody or the way to speak can be described by a space in five dimensions: the intonation, the intensity, the speech rate, the vocal quality and the degree articulation [22]. If most of these dimensions are today measurable in batch mode [23], some of them are also measurable in real time. It is the case of the intonation (yin [24]) and of the intensity (loudness). We added a speech rate estimator (syllex). These three real-time speech descriptors allow to inform us about the way the performers utter. They can be used, in the same way as the sensors, to control the various audio engines. If the modification of the intonation by transposition and that of the intensity by variable gain are henceforth known and well enough mastered in real time, it does not also go away for the speech rate. Now we expose in this section, one new paradigm allowing the transformation of the speech rate in real time. All the prosodic transformations used are available in the SuperVP [25] audio engine, the IRCAMs high quality phase vocoder in real time . In fact, the SuperVP library already implements a quality transposition, as well as some other modifications such as the spectral envelope transformation. One of the objects of this library, SuperVP.ring, allows to make these modifications on a circular buffer the size of which can be arbitrarily defined (3 seconds in our case). The advantage of the circular buffer is to keep the instantaneousness of the transformation, while enabling, at any time, to be able to move in short-term past (term equivalent to the size of the buffer). Thanks to it, we can locally stretch out certain portions of the signal as the vowels (using a real time voicing detection) and provoke at the listener’s the perception of a slowing down of the speech rate. If we cannot move to the future, the return in the immediate position can be made in a accelerated way, provoking, this time, the perception of an acceleration of the speech rate. As if the read head of the buffer behaved as an elastic which we stretch out and relax. Extremely, it is possible to freeze the read head of the buffer in a place that provokes a “stop on sound” who can give interesting effects (extremely long vowels for example that makes speech sounds like sing).

4. EXAMPLES OF MAPPING

In this part, we give some examples of mapping between the control data and the parameters of the audio engines. The figure 3 lists the various available controllers (to the left), as well as the various parameters of the audio engines (to the right). The mapping consists in connecting the controllers with the parameters (by some linear or non-linear scales) and it can vary in the time, as it is the case in *Luna Park*. Two types of connections are possible: the discrete connections and the continuous connections. Indeed, the discrete controllers (underlined on the figure 3), giving only a value from time to time, as the hit energy

estimator, correspond to the control of type percussive and are going to serve for controlling the discrete parameters of the audio engines, as the activation of a segment for the concatenative synthesis (highest arrow). On the contrary, a continuous connection connects a continuous controller, as the linear absolute position on a tactile ribbon in a continuous parameter of audio engines such as the transposition, for instance (lowest arrow).

Some scenarios chosen for the piece are described be-

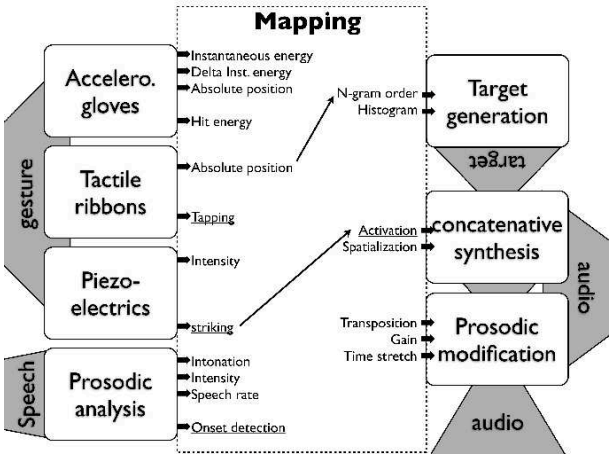

Figure 3. The mapping is at the interface between the data stemming from controllers (to the left) and the parameters of the audio engines (to the right). As example, two arrows were drawn. The highest allows to make vary the order of N-gram using a tactile ribbon. The lowest allows to activate a segment of the concatenative synthesis by a striking on the body.

low. By connecting in a direct way the hit energy estimator of the right glove of the percussionist (who is right-handed) with the activation of the synthesis, he can manage the speech rate and segmental aspects by percussive movements of the right hand. If the rotation of the left hand is connected with the transposition and with the intensity of the synthesis, he can then control the prosody of this one with both hands. In a scene of the piece, a dancer caresses a tactile ribbon situated over a screen transverse. The speed of her caress, deduced from the detected absolute linear position of her finger, is then connected with the speech rate modification of another performer's speech, which is speaking at the same moment. Thanks to a mapping favoring the catch of the vowels, she can manage to make "a stop on sound" in a vowel which gives the effect that she hold his tong. The accumulation of the energy of the hands of the percussionist is used, in a scenario, to control the N-gram order of the generation of targets. The activation of the segments is here automatic (continuous stream where every end of segment activates the next one) and controller changes only the order in which segments are chosen. The more the percussionist is energetic and the more the order of N-gram decreases, going to the random for large-scale movements.

5. FUTURE PERSPECTIVES

Of this period of research can be deduced several perspectives concerning the gesture capture, the real time speech synthesis, as well as their connections. Concerning the gesture capture, the used sensors possess the advantage to be rather light to be integrated into gloves, as well as good sensitivity in the movement (capture of the variation of the momentum is accurate). On the other hand, they present a rather low energy autonomy, and do not offer the measure of the absolute static position. It would be beneficial to add to accelerometers, another technology allowing to access the absolute position. As regard the audio engine, it would be interesting to bend over the other speech synthesizers based on parametric (articulatory), semi-parametric (HMM) or hybrid models (concatenative/HMM). Indeed, the concatenative synthesis in batch mode has for advantage its degree of realism, which becomes difficult to maintain in real time. Finally the mapping between the gesture and the speech synthesis is a rich subject of research. As a research track, we can imagine more complex mappings where become interleaved the temporal and the semantic aspects of both the hand gesture and the vocal gesture.

6. ACKNOWLEDGMENTS

Author would like to thank Fred Bevilacqua, Bruno Zamborlin, Norbert Schnell, Diemo Schwarz, Riccardo Borghesi, Emmanuel Fléty, Maxime Le Saux, Pascal Bondu, Xavier Rodet, Christophe Veaux, Pierre Lanchantin and Jonathan Chronic, for their helps.

7. REFERENCES

[1] S. Lemouton, "Utilisation musicale de dispositifs de captation du mouvement de l'archet dans quelques oeuvres récentes," in *JIM*, 2009.

[2] F. Bevilacqua, N. H. Rasamimanana, E. Fléty, S. Lemouton, and F. Baschet, "The augmented violin project: research, composition and performance report." in *NIME*, 2006.

[3] F. Bevilacqua, F. Guédy, N. Schnell, E. Fléty, and N. Leroy, "Wireless sensor interface and gesture-follower for music pedagogy," in *NIME*, 2007, pp. 124–129.

[4] P. Cook, "Real-time performance controllers for synthesized singing," in *NIME*, 2000.

[5] C. d'Alessandro, N. D'Alessandro, S. L. Beux, J. Simko, F. Cetin, and H. Pirker, "The speech conductor: gestural control of speech synthesis," in *eNTERFACE*, 2005.

[6] N. D'Alessandro, C. d'Alessandro, S. L. Beux, and B. Doval, "Real-time calm synthesizer: New approaches in hands-controlled voice synthesis," in *NIME*, 2006, pp. 266–271.

[7] S. Fels and G. Hinton, “Glove-talk 2: A neural network interface which maps gestures to parallel formant speech synthesizer controls,” *IEEE Transactions on Neural Networks*, vol. 9, no. 1, pp. 205–212, 2004.

[8] P. Cook and C. Leider, “Squeeze vox: A new controller for vocal synthesis models,” in *ICMC*, 2000.

[9] P. Cook, “Spasm: a real-time vocal tract physical model editor/controller and singer: The companion software synthesis system,” *Computer Music Journal*, vol. 17, no. 1, pp. 30–34, 1992.

[10] T. Hueber, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, “Towards a segmental vocoder driven by ultrasound and optical images of the tongue and lips,” in *Interspeech*, 2008, pp. 2028–2031.

[11] B. Denby and M. Stone, “Speech synthesis from real time ultrasound images of the tongue,” in *ICASSP*, 2004, pp. 685–688.

[12] A. Esposito, M. Faundez-Zanuy, E. Keller, M. Marinaro, B. Kröger, and P. Birkholz, “A gesture-based concept for speech movement control in articulatory speech synthesis,” *Verbal and Nonverbal Communication Behaviours*, no. 4775, pp. 174–189, 2007.

[13] E. Fléty and C. Maestracci, “Latency improvement in sensor wireless transmission using ieee 802.15.4,” in *NIME*, 2011.

[14] N. Schnell, R. Borghesi, D. Schwarz, F. Bevilacqua, and R. Müller, “FTM-Complex Data Structures for Max,” in *ICMC*, Barcelona, Spain, Sep. 2005.

[15] N. Schnell, A. Röbel, D. Schwarz, G. Peeters, and R. Borghesi, “Mubu and friends - assembling tools for content based real-time interactive audio processing in max/msp,” in *ICMC*, 2009.

[16] D. Schwarz, G. Beller, B. Verbrugghe, and S. Britton, “Real-time corpus-based concatenative synthesis with catart,” in *DAFx*, 2006.

[17] G. Beller, D. Schwarz, T. Hueber, and X. Rodet, “Hybrid concatenative synthesis in the intersection of speech and music,” in *JIM*, A. Sedes and H. Vaggione, Eds., vol. 12, 2005, pp. 41–45.

[18] C. Veaux, G. Beller, and X. Rodet, “Ircamcorpustools: an extensible platform for spoken corpora exploitation,” in *LREC*, Marrakech, Morocco, may 2008.

[19] G. Beller, C. Veaux, G. Degottex, N. Obin, P. Lanchantin, and X. Rodet, “Ircam corpus tools: Système de gestion de corpus de parole,” *TAL*, 2009.

[20] P. Lanchantin, A. C. Morris, X. Rodet, and C. Veaux, “Automatic phoneme segmentation with relaxed textual constraints,” in *LREC2008*, Marrakech, Morocco, 2008.

[21] G. Beller, “Transformation of expressivity in speech,” *Linguistic Insights*, vol. 97, pp. 259–284, 2009.

[22] ——, “Analyse et modèle génératif de l’expressivité : Application à la parole et à l’interprétation musicale,” Ph.D. dissertation, Université Paris XI, IRCAM, June 2009.

[23] G. Beller, D. Schwarz, T. Hueber, and X. Rodet, “Speech rates in french expressive speech,” in *Speech Prosody 2006*, SproSig. Dresden: ISCA, 2006, pp. 672–675.

[24] A. D. Cheveigné and H. Kawahara, “Yin, a fundamental frequency estimator for speech and music,” *JASA*, vol. 111, pp. 1917–1930, 2002.

[25] A. Roebel, F. Villavicencio, and X. Rodet, “On cepstral and all-pole based spectral envelope modeling with unknown model order,” in *PRL*, 2006.